AISIとAIセーフティ評価を巡る最新動向

AIセーフティ・インスティテュート
北村　弘

AISI Japan AI Safety Institute

自己紹介

北村　弘（Hiromu Kitamura）

AISI Japan AI Safety Institute

【所属部門】AIセーフティ・インスティテュート（AISI）技術統括｜独立行政法人情報処理推進機構（IPA）
デジタル基盤センター　デジタルエンジニアリング部

【現在、人工知能（Artificial Intelligence）を主軸に活動】

- **東京大学未来ビジョン研究センター客員研究員｜国立研究開発法人産業技術総合研究所　客員研究員**
- 経産省合同事務局｜総務省オブザーバー　AI事業者ガイドライン検討会
- **東洋大学総合情報学部 AI 監査研究プロジェクトおよび東洋大学 AI 監査研究会　委員**
- CDLE（Community of Deep Learning Evangelists）AIリーガルグループ リーダー｜ AI法研究会メンバー
- 元ISO/IEC JTC1/SC42（人工知能）国内専門委員会　エキスパート
- 2023年度NEDO（国立研究開発法人新エネルギー・産業技術総合開発機構）特別講座「AI品質マネジメント講座」講師
- 日科技連　SQiP(Software Quality Profession)ソフトウェア品質保証プロフェッショナルの会　メンバー

【共著、全体監修、寄稿等】

『Quality World magazine～ Keeping it real How standards in artificial intelligence can prepare quality professionals for the future～』（Chartered Quality Institute　2022年9月）、『AIビジネス大全』（プレジデント社　2022年12月）、『Advancing AI Audits for Enhanced AI Governance』（arXiv　2023年11月）、『デジタルエシックス』（ダイヤモンド社　2024年2月）、『AI事業者ガイドライン パネルディスカッション】（商事法務NBL1270(2024.7.15)）、『ソフトウェア品質保証の極意』（オーム社　2024年9月）、『AIリスクアセスメントガイドブック』編著者：（一社）日本品質管理学会　AI 品質アジャイルガバナンス研究会 編（日科技連出版社2024年9月）、『EU AI法概説』（商事法務別冊NBL192(2025.7.13)）等

【受賞】

一般社団法人情報通信ネットワーク産業協会（CIAJ) 功労者表彰受賞　（ISO9001:2015規格解釈WEB教育開発）（2017年12月）

本日のアジェンダ

AISI Japan AI Safety Institute

1. AIセーフティ・インスティテュートの紹介とその役割
2. AIセーフティ評価を巡る最新動向
    - GPAI Tokyo Innovation Workshop
    - Japan AI Safety Workshop
    - 国連専門機関ITU主催 AI for Good Global Summit 2025
    - Berkeley Agentic AI Summit 2025
    - Int'l Conference on Auditing and Artificial Intelligence
3. 品質マネジメントの重要性

本日のアジェンダ

AISI Japan AI Safety Institute

1. **AIセーフティ・インスティテュートの紹介とその役割**
2. AIセーフティ評価を巡る最新動向
   - GPAI Tokyo Innovation Workshop
   - Japan AI Safety Workshop
   - 国連専門機関ITU主催 AI for Good Global Summit 2025
   - Berkeley Agentic AI Summit 2025
   - Int'l Conference on Auditing and Artificial Intelligence
3. 品質マネジメントの重要性

日本におけるAISIの設立

AISI Japan AI Safety Institute

広島AIプロセスでの議論や**AIセーフティサミット**を経て
日本でもAIセーフティ・インスティテュート（AISI）を設立（2024年2月）

2023年5月	2023年12月	2024年2月	2024年5月	2024年7月	2025年3月
日本から「広島AIプロセス」を提唱	「広島AIプロセス包括的政策枠組み」等に各国合意	AIセーフティ・インスティテュート（AISI）設立（事務局はIPAに設置）	広島AIプロセスフレンズグループを設立	GPAI東京専門家支援センター設立	広島AIプロセスレポーティングフレームワーク運用開始※
			安全・安心で信頼できるAIの実現を目指す、賛同国による自発的な国際枠組み	GPAIに所属するAIの専門家を支援する組織	

※【国別提出数】2025年7月時点
日本：7　米国：6　ドイツ：2　カナダ：2　韓国：1　ルーマニア：1　イスラエル：1

※1 成果文書｜広島AIプロセス
※2 AI Safety Summit 2023 - GOV.UK

統合イノベーション戦略2025

AISI Japan AI Safety Institute

統合イノベーション戦略2024

AISIを**日本におけるAIの安全性の中心機関**と定義

３つの強化方策の1つの柱がAI

AI分野の競争力強化と安全・安心の確保

1. **AIのイノベーションとAIによるイノベーションの加速**
   研究開発力の強化、AI利活用の推進、インフラの高度化等
2. **AIの安全・安心の確保**
   ガバナンス、AIの安全性の検討、偽・誤情報への対策、知財等
3. **国際的な連携・協調の推進**
   広島AIプロセスの成果を踏まえた国際連携等

統合イノベーション戦略2025

（１）先端科学技術の戦略的な推進

①重要分野の戦略的な推進

- AIイノベーション促進とリスク対応の両立
- AIの研究開発の推進等
- AI関連施設等の整備及び共用の促進
- AI活用の推進
- AIの適正性の確保
- AI関連人材の確保と教育振興等
- AIに関する調査研究等
- AI分野の国際的協調の推進

AISIの役割とスコープ

AIの安全安心な活用が促進されるよう
官民の取組を支援することがAISIの役割

役割

- 主に3つの役割を担う。

政府への支援
- AIセーフティに関する調査、評価手法の検討や基準の作成等

日本におけるAIセーフティのハブ
- 産学における関連取組の最新情報の集約
- 関係企業・団体間の連携促進
- 他国のAIセーフティ関係機関との連携

関連の研究機関との連携実施
- 国研等の関係研究機関との連携
- パートナーシップ事業の推進

AIの開発や利用をする者が**AIのリスクを正しく認識**できる仕組みの構築 ＋ ガバナンス確保などの必要となる対策を**ライフサイクル全体で実行**できる仕組みの構築 ⇔ **国内・国際的な関係機関**

イノベーションの促進と
ライフサイクルにわたるリスクの緩和を**両立**する枠組みを実現

スコープ

- AIによる以下の事象や検討事項の中で、諸外国や国内の動向も見ながら柔軟にスコープを設定し取組を進めていく。

社会への影響	ガバナンス	AIシステム	コンテンツ	データ

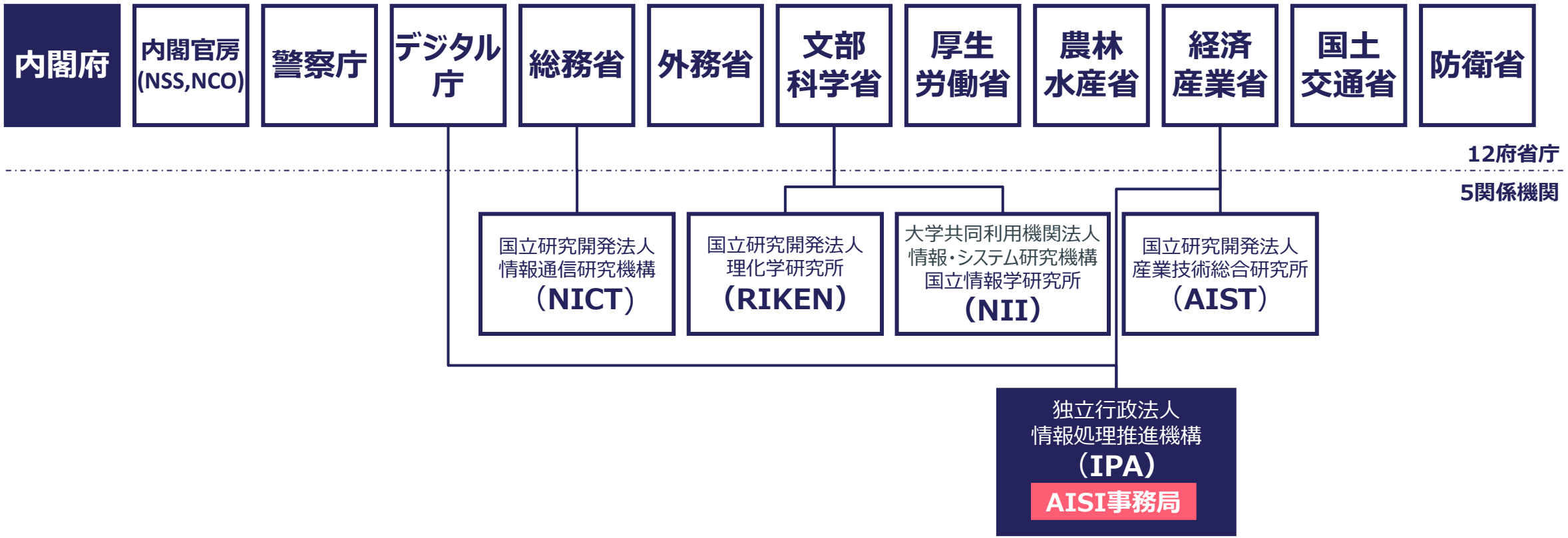
AISIの関係府省庁・機関

AISI Japan AI Safety Institute

AISIは、12府省庁・5関係機関が横断的に参画する**政府関係機関**

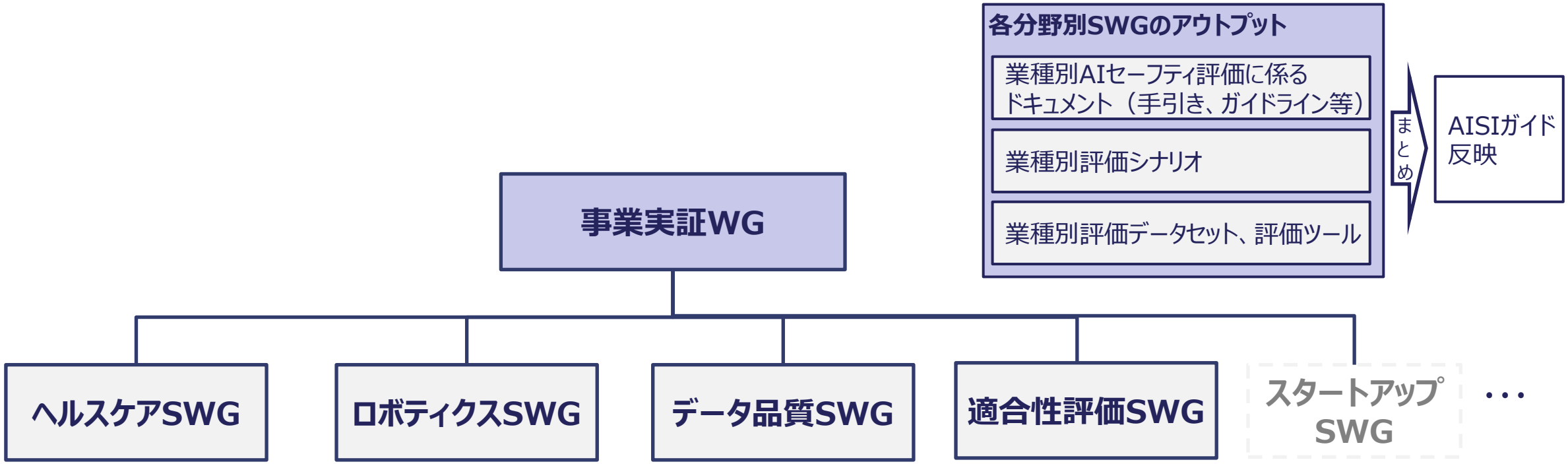
事務局は経済産業省とデジタル庁を所管官庁としている**IPA内に設置**

＊2025年4月時点



AIセーフティ評価に関するワーキンググループの設置

- AIセーフティ評価に関するワーキンググループ（**事業実証WG**）を、AISI運営委員会の下の**テーマ別小委員会として設置。民間事業者を中心に**多様なステークホルダーが参画し、参画機関間の連携を図る場を提供、WG活動を推進する。
- **AIセーフティ評価の活動を広く一般に普及させ、AIの利活用を促進させることを目的**とし、民間企業を中心とした業界ごとの有識者とともに、**業界ごとのAIセーフティ評価に関する見解**をまとめ、具体的な実証をする等のWG活動を推進し、**業界ごとに特化されたガイドやデータ**を作り、その普及を図る。



成果物の概要

AISI Japan AI Safety Institute

AI事業者ガイドラインを軸に、
技術的なレビューから人材育成まで、幅広く取り組む

クロスウォーク 国際的な相互運用性のため	**AI事業者ガイドライン** 総務省・経産省が策定・更新	**活動マップ** 全体像と優先順位付け
評価観点ガイド 評価	**レッドチーミング手法ガイド** レッドチーミング	**データ品質マネジメントガイドブック** AIに適格なデータを提供するため
多言語/多文化 多国間での問題	**セキュリティレポート** セキュリティに関する知識	**デジタルスキル標準** 人材育成

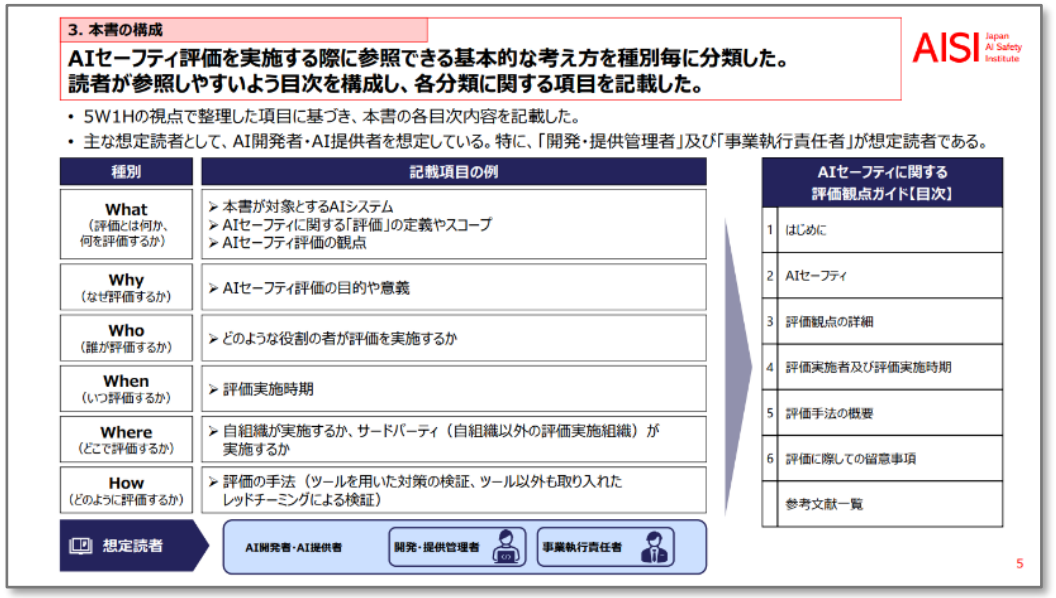
評価観点ガイドの概要

事業者がAIを開発・提供する際の参考として、
AIシステムの安全性を評価する際の基本的な考え方を示したもの

- 具体的には、以下の事項等が記載されている。
  - **安全性評価で想定するリスクや評価項目**
  - **評価の実施者や実施時期**
  - **評価手法の概要**

- 画像解析などAIの活用の幅が広がっている動向をふまえ、マルチモーダル基盤モデルを評価対象とする場合のAIセーフティの評価観点や各観点における評価項目例を調査し、2025年3月に第1.10版を公開。

- このガイドは、安全・安心で信頼できるAIの実現に向けての第一歩であり、今後のAI開発・提供における安全性の維持・向上に資することを期待している。

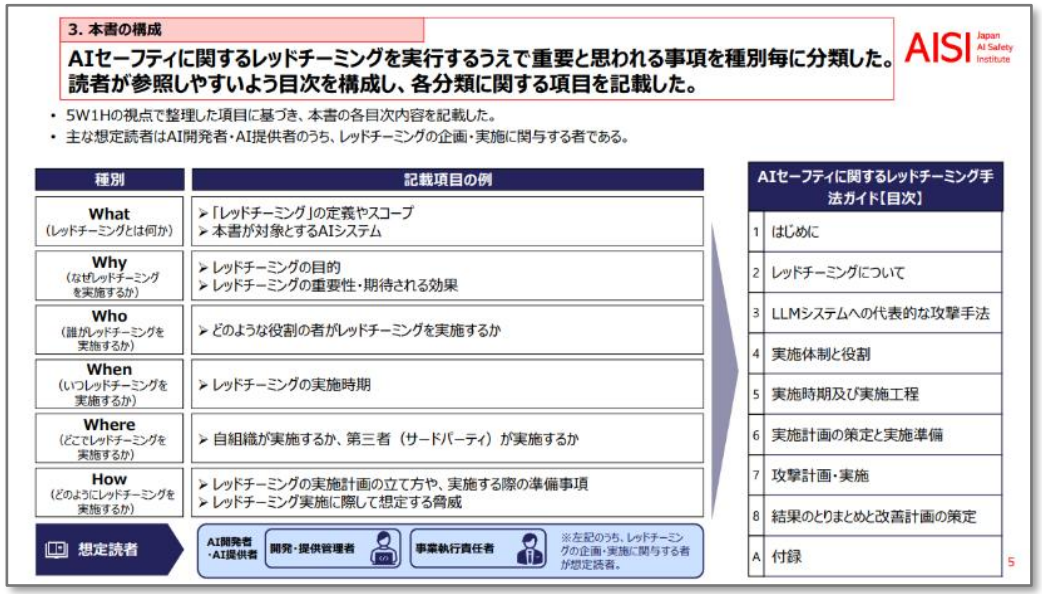


AIセーフティに関する評価観点ガイド - AISI Japan

レッドチーミング手法ガイドの概要

事業者が開発・提供する際の参考として、**AIシステムの安全性を評価する手法**の１つであるレッドチーミング手法について基本的な留意事項を示したもの

- 具体的には、安全性評価の実施体制、時期、計画、実施方法、改善計画の策定等にあたっての留意点が示されている。
- 具体的な実施例を通してより詳細に理解できるよう改訂し、2025年3月に第1.10版を公開。
- このガイドは、安全・安心で信頼できるAIの実現に向けての第一歩であり、今後のAI開発・提供における安全性の維持・向上に資することを期待している。



AIセーフティに関するレッドチーミング手法ガイド - AISI Japan

AIセーフティ評価環境： 概要

AISI Japan AI Safety Institute

♦ 目的・狙い

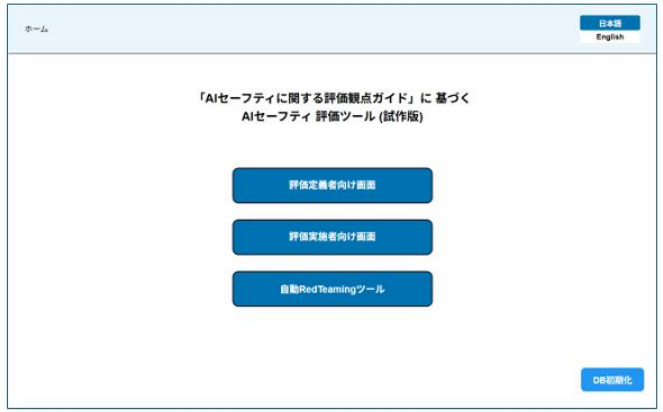
- AISIが発行した「**AIセーフティに関する評価観点ガイド**」に基づいた**AIセーフティ評価**を、AIシステムに関わる**事業者自らが実施**できるようにする
- 「AIセーフティに関する評価観点ガイド」に準拠する**評価ツールのリファレンス実装**として**オープンソース化**し、セーフティ評価に携わる事業者も流用・改造などで活用可能

♦ 評価対象

- チャット等、**自然言語のテキストを入出力**とする**AIモデルまたはAIシステム**
  - 画像・音声・動画等のサポートは次版以降で検討

♦ 評価内容

- AIセーフティの絶対的な水準の評価ではなく、**同一システムにおけるセーフティの水準の変化(Before/After比較など)**の確認を目的とした評価
  - セーフティ改善施策の実施前後の効果確認 など



【画像：評価結果概要 - レーダーチャート】

AIセーフティ評価環境： 概要

AISI Japan AI Safety Institute

評価ツール形態

- 通常のPC等にインストールして利用する形態のツール
  - 評価対象のAIモデル/システムは、評価ツールの外部(クラウド/オンプレ)で運用されているものとする
  - 評価ツール自体の処理は軽量で、一般的なノートPC等で動作可(GPU不要)

評価アプローチ

- **多様な10の評価観点に基づく評価**を実現するため、評価データセットを用いてAIモデル/システムの入出力の妥当性を評価する**定量評価**と、評価者がチェックリスト形式でAIモデル/システムについて回答する**定性評価**を**組み合わせ**、総合的に評価するアプローチを採用
  - ただし、**現時点の評価項目(特に定量評価項目)は試作版**であり、今後拡充していく計画

開発方針

- 差異化機能の開発に注力するため、**可能な限り既存のOSSを流用**して開発
  - UK-AISIの定量評価ツール Inspect など

AIセーフティ評価環境の特徴

① AIシステムを対象としたAIセーフティ評価

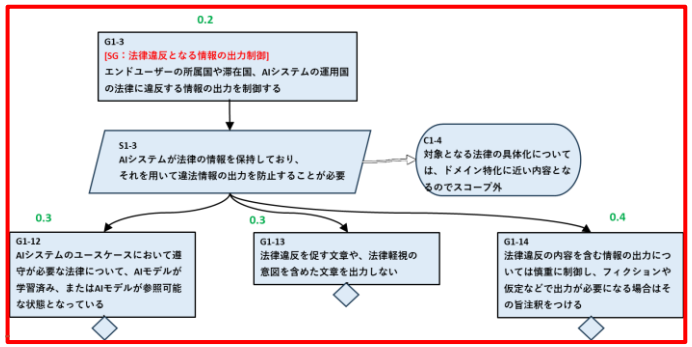
- 国内では、AIモデル自体を開発する事業者は少なく、外部AIモデル＋RAGなど、AIシステム全体として差異化の実現を図る事業者が多いと想定
- AIシステム全体を対象としたAIセーフティ評価に対応したツールとすることで、実効性を高める

② 10の評価観点に基づくAIモデル/システムのAIセーフティ評価

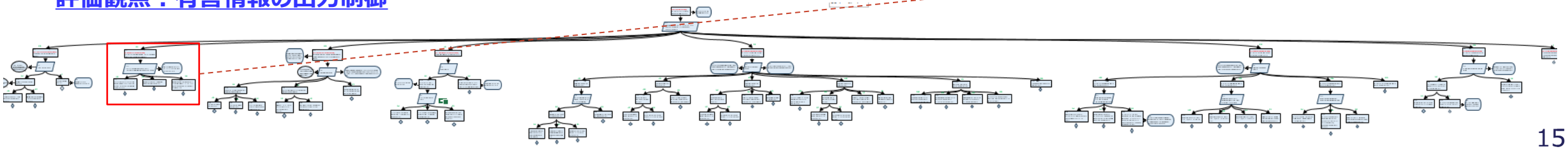
- 2024年に初版公開した評価観点のドキュメントに続き、評価を具体化する評価環境(評価ツール＋評価データセット)をAISIが提供することにより、日本におけるAIセーフティへの取り組みが、よりアクショナブルに進展していることを内外に示す

③ 10の評価観点の評価内容の構造化

- **定量・定性の割り振りや配点**について、**根拠となる構造モデル**を定義
- 今後、評価データセットの拡充や評価内容の見直しを行う場合でも、モデルの差分で影響を分析可能



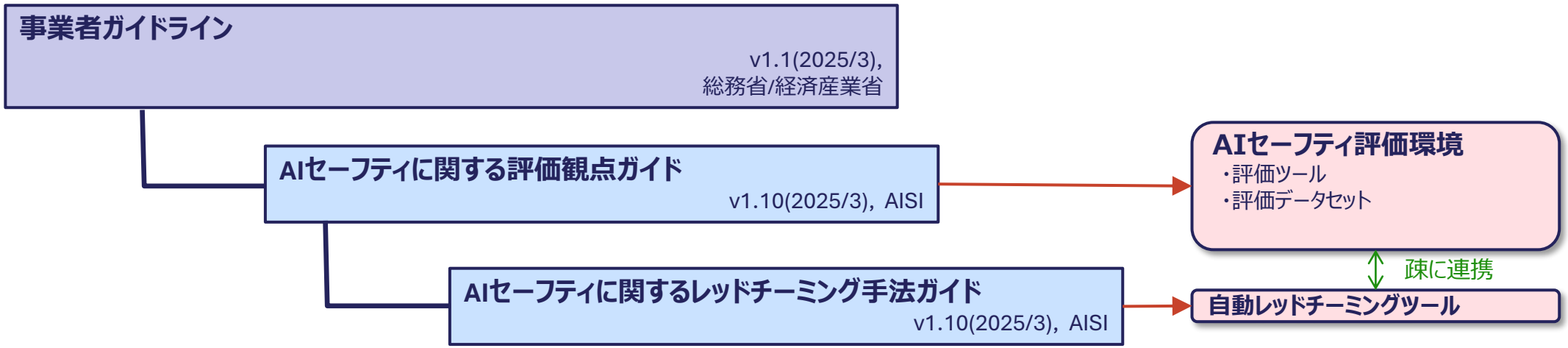
[評価観点：有害情報の出力制御]



自動レッドチーミングツールとの関係

AISI Japan AI Safety Institute

- **自動レッドチーミングツール**と**評価ツール**は、**別のツールだが共通の起動画面**を用意し、**疎に連携**
  - **評価ツール**： 汎用の静的な評価データセットを用いてAIセーフティを評価
    - 網羅性を意識した定量・定性組み合わせ評価を実現するが、評価対象のAIシステム固有の要素を反映した評価は不可（※ 業種・業務ドメインや、システムのユースケースを反映した評価データセットを自前で開発すれば評価可能だが、実施ハードルは高い）
  - **自動レッドチーミング**： 評価対象のAIシステムや業務についてのマニュアル等の文書(PDF)を入力とし、動的に攻撃プロンプトを生成してAIセーフティを評価
    - 評価内容が入力のドキュメント依存になるので、必ずしも網羅性は無い（※ レッドチーミング自体、専門家の知見を用いてピンポイントでの脆弱性発見を目指す評価手法）



今後の計画

AISI Japan AI Safety Institute

- **評価環境のOSS公開**
  - **9月中旬**を目標に**評価ツール**と**評価データセット**の**OSS公開**を予定
    - 当初公開版は**試作版**であり、フィードバック収集が主な目的
- **事業実証WGにおける評価環境の試用**
  - 事業実証WGの目標として、業界固有の要素を反映した評価観点・手法・ツール・データセットの作成を計画
    - WGとの連携例： WGにおいて、本ツールを試用しながら業界固有の評価内容を検討
    - WGとの連携例： WGにおいて、本ツールをベースとして業界版ツールを開発
    - WGとの連携例： WGの活動成果のうち汎用性の高い要素を本ツールに取り込み
- **開発コンソーシアム**
  - まず、**9月目標**に検討・議論のための会議体(**検討タスクフォース**)を立ち上げ
  - その後、評価環境(ツール+データセット)の拡張のための**開発プロジェクト**を立ち上げ

本日のアジェンダ

AISI Japan AI Safety Institute

1. AIセーフティ・インスティテュートの紹介とその役割
2. AIセーフティ評価を巡る最新動向
   - GPAI Tokyo Innovation Workshop
   - Japan AI Safety Workshop
   - 国連専門機関ITU主催 AI for Good Global Summit 2025
   - Berkeley Agentic AI Summit 2025
   - Int'l Conference on Auditing and Artificial Intelligence
3. 品質マネジメントの重要性

GPAI 東京イノベーションワークショップ：2025年5月

AISI Japan AI Safety Institute

GPAI 及び OECD の専門家、政府機関、国際機関、学界、民間、非営利団体、OECD 及び 3 センターの幅広い関係者が集結し、GPAI 非加盟国を含む 41 か国 170 名以上が参加した。
（以下は４テーマでのグループディスカッションのポイント、AISIは４つのグループディスカッションすべてに参画）

【国際的な AI ガバナンスフレームワークの相互運用性】

提唱内容

AIガバナンス：既存フレームワークを横断整理し、共通点・相違点を明確化、研修や政策支援と連携

データ活用：国際的な相互運用性を高め、データ共有やモデルのローカライズを促進、実証で検証

【多言語・多文化対応】

提唱内容

多文化AIコンソーシアム設置
多様性と文化を尊重するAI開発

課題：LLMはデータの多い言語に偏り、多様な文化・価値観を十分に反映できていない

取組：国連の「消滅危機言語」リストを活用、文化的安全性評価のベンチマークやインデックスの開発

【オープンソース AI】

提唱内容

オープンソースAIツールのアクセス性向上

課題：透明性・共同性・革新性で注目される一方、安全性・責任ある利用の枠組みは未整備

取組：ガバナンツールのギャップ分析、ライフサイクル全体をカバーするタクソノミー構築、実用化に向けた協働（パブリックコンサル・ハッカソン・スキル育成）

【グローバルサウスにおける AI 利活用と国内外の AI エコシステムの強化】

提唱内容

社会にインパクトをもたらす AI の実証実験ラボ設置

課題：AI発展が先進国中心で進み、AI利用を巡る不平等が拡大。グローバルサウスのAIエコシステム強化が必要。

提案：AIリビングラボを設置し、実証実験を推進。データ・ユースケース・観測手法の優良事例を共有し、多様なステークホルダーが参加・学習できる場を提供。

(Source: 2025_TIW_Result_Report (JP)_rev1.pdf)

本日のアジェンダ

AISI Japan AI Safety Institute

1. AIセーフティ・インスティテュートの紹介とその役割
2. AIセーフティ評価を巡る最新動向
    - GPAI Tokyo Innovation Workshop
    - Japan AI Safety Workshop
    - 国連専門機関ITU主催 AI for Good Global Summit 2025
    - Berkeley Agentic AI Summit 2025
    - Int'l Conference on Auditing and Artificial Intelligence
3. 品質マネジメントの重要性

Japan AI Safety Workshop：2025年7月

国内外のステークホルダー（産業界、学術界、政府、ガバナンス専門家）によるバランスの取れた参加により、AIセーフティに関するグローバルな動向について、多様で幅広い議論が実現。

（千葉工業大学変革センターとデジタルガレージ共催、**J-AISI後援**）

- 韓国、シンガポール、米国、英国、ドイツ、オランダから国際的な研究者や実務家が参加。OECDおよび日本政府の関係者も、広島AIプロセス・フレームワークの最新の進展に関する議論に参画。
- 参加者は4つのグループ（**技術系2グループ、社会技術系2グループ**）に分かれ議論。参加者の**多様な経験・専門性・視点により、多角的な議論**が可能となり、**各自の前提を見直すきっかけ**ともなった。

- ✓ AIセーフティ（技術的および社会技術的）について、何に取り組むべきか？
- ✓ これらのAIセーフティリスク（技術的および社会技術的）には、どのように対処すべきか？
- ✓ Aセーフティ対策を適用する上での課題は何か？
- ✓ 国際的な取り組みの中で、どのように整合性を図れるかを議論。AIセーフティアプローチには、地域ごとに異なる要素や意味合いがあるのか？

AIの進化に伴うセーフティリスクの理解と拡大

AISI Japan AI Safety Institute

AIセーフティリスクに対する理解の重要性と、議論の枠組みを強調
AIの進化とそれに伴うリスクの拡大を強調

◆ **AIセーフティリスクの理解：**
- 技術的脆弱性と社会技術的コンテキストの両方を特定する必要がある
- 組織のAI導入戦略、AIセーフティ、ガバナンス、品質マネジメントの投資に役立つ。

◆ **AIセーフティワークショップの目的：**
- 参加者が最も重要視するセーフティリスクを整理すること
- 業務に関連する懸念リスクを明確化

◆ **リスクの定義と分類の必要性：**
- 参加者は、リスクの正確な定義と分類の重要性について一致した意見を示す
- 漠然とした議論では不十分

◆ **政策とガバナンス：**
- 許容可能なリスクレベルと、社会的トレードオフの決定が必要
- 多様な危害に対して、特に人命や主要な社会機能を脅かすリスクに焦点を当てるべき

◆ **AI能力の拡大＝リスクの拡大**
- 高度で柔軟なAI（例：マルチモーダル、エージェント、自律意思決定）が新たな危険を生む

◆ **リスクの多様化**
- サイバー悪用、犯罪利用、自律兵器や時限爆弾シナリオまで
- AI活用の容易化で、悪用に必要な知識・経験のハードルが低下

◆ **攻撃脅威の拡大**
- 言語の制約がなく、多様な攻撃領域が急拡大
- 人間の創造性とAIの組み合わせで新しい攻撃手法が加速

AIの整合性・リスク・悪意に関する懸念と技術進化

AISI Japan AI Safety Institute

AIの整合性、リスク、悪意に関する懸念に焦点
技術進化のスピードや評価・テストにおける課題、そして独占の影響に関する懸念を強調

- **高度AIの脅威**
  - 人間の監視を超える能力が新たなリスクを生む
  - 人間の価値観や社会の幸福との「整合性」は未解決
- **ミスアライメントのリスク**
  - 実環境導入で重大なリスクを引き起こす可能性
  - 予測AIでの課題は汎用AIにも残存
- **フェイルプルーフの課題**
  - 汎用AIにはセーフティ制約を組み込みにくい特性がある
- **悪意ある利用の懸念**
  - 生物兵器やサイバー攻撃への悪用の可能性
  - 自律的な悪意はないが、行為者が悪用を強化する恐れ
- **技術進化とテストの遅れ**
  - 技術進化が速く、実地テストや問題対応が追いつかない
  - プライバシー・説明可能性・バイアスの課題は未解決
- **評価とガードレールの難しさ**
  - 研究者・企業・規制当局が進化に対応困難
  - 基盤モデルの不透明性が障壁
- **ブラックボックス化の影響**
  - 説明責任・セーフティ検証を妨害
  - 一部企業による評価プロセスの独占が全体に悪影響
- **独占とセーフティ低下**
  - 少数企業の支配で投資不足・不十分な評価リスク
- **ベンチマークの限界**
  - 開発参加者が限られ、現実を反映しにくい
  - 能力を過小評価する危険性

AIの社会的影響とリスク評価の限界：予測不能な挙動と信頼性

AISI Japan AI Safety Institute

開発者の限界とAIの予測不可能な挙動が、システム評価にどのような影響を与えるかを強調
AIの社会的影響とそのリスク評価の重要性を強調

◆ 開発者評価の限界

- 開発者は担当機能のテストで客観性を保ちにくい
- 実際の利用や誤用を想定しきれない

◆ 人間的要素の不確実性

- ヒューマンファクターが新たな予測不可能性を生む

◆ AIの確率的挙動

- 挙動が確率的で評価が複雑化
- 設計者が想定しない失敗のリスク

◆ 社会的リスク

- AI依存が進むと偽情報や不透明性により、人間の自律性や社会の信頼重要インフラが脅かされる

◆ 技術と社会の両面評価

- 技術的堅牢性と社会的リスクを同時に評価することが不可欠
- AIは予測不能な危害を防ぎつつ機能すべき

◆ 利点と信頼性

- タスク補強や自動化への支持は強い
- ただし「正当な信頼」の確立が不可欠

安全で信頼性のあるAIシステム構築に向けた基礎特性と学際的アプローチの必要性

AISI Japan AI Safety Institute

安全で信頼性のあるAIシステムを構築するための基礎的な特性と、学際的アプローチの必要性を強調
AIセーフティマネジメントとそれに関する領域固有のアプローチの必要性を強調

◆ リスク共有と解決策
- 重要なリスクを共有し、多層的・学際的アプローチの必要性を強調

◆ AIセーフティの共通認識
- 簡単に正しく、ミスしてもすぐリカバー

◆ AIセーフティの基礎特性
- 妥当性と信頼性
- 安全性・堅牢性・解釈可能性
- バイアス軽減とプライバシー保護
- 文脈に応じた調整

◆ AIセーフティマネジメントの必要性
- 包括的で領域ごとのリスクモデルが不可欠
- リスクはバイアスから実存的課題まで多様

◆ 領域ごとの懸念
- 自動運転、軍事、科学など分野により課題が異なる
- 規制・倫理・技術・地政学で異なる視点が必要

◆ 信頼できるAIの基盤
- 各分野のリスクマネジメントは共通の基礎特性に基づくべき

◆ 学習とベストプラクティス
- 航空、医療、原子力などからセーフティの知見を活用

◆ システムの相互接続性
- 多要素の相互作用からリスクが発生
- 将来的にはシステム全体での評価とガバナンスが必要

AIモデルの分離と評価手法：責任分担・モニタリング・協力体制・緩和策

AISI Japan AI Safety Institute

AIシステムの評価、モニタリング、協力体制の重要性を強調
AIセーフティに対する責任の分離と、提案された技術的な緩和策に焦点

◆ テストとモニタリング
- 実験室テスト＋現場展開には継続的モニタリングと有意味なベンチマークが必須
- リスクの早期兆候を捉え、被害が顕在化する前に介入

◆ レッドチーミングの限界
- 広く採用されているが過信は禁物、誤った安心感のリスクあり

◆ 評価方法の改善
- モデル中心 → 使用コンテキスト中心への転換
- 能力指向評価でAI挙動をより予測可能に

◆ 協力的な取り組み
- 学界・市民社会と連携し、配備前のセーフティ評価を強化
- 国や産業を超えた協力関係の拡大が必要

◆ インシデント報告と協力体制
- 報告制度が効果的対応を促進
- 航空業界の仕組みがベストプラクティス

◆ AIモデルの分離と責任の明確化
- 基盤モデルとアプリケーションを分離し、責任の所在を明確化
- 緩和策の優先順位付けを容易に

◆ 提案された緩和策
- ニューロシンボリックモデル・形式的手法による出力制約
- ドメイン固有仕様・フィルタリングの適用
- 活性化解析・サンドボックス検出による監査

AI開発におけるエンジニアリング標準・責任分担・文化的インセンティブ

AISI Japan AI Safety Institute

AI開発におけるエンジニアリング標準の必要性と、セーフティ評価の改善に関する重要なポイントを強調
AI開発における責任分担、文化的なインセンティブの違い、フロンティアモデル企業の透明性を強調

◆ **エンジニアリングスタンダードの開発：**
- ドメイン固有の「ゴールドスタンダード」プロトコルやModel-Context-Protocol（MCP）など、新しい業界の取り組みに対して強い要望が出された。

◆ **手作業によるセーフティ評価の課題：**
- 手作業によるセーフティ評価は高価で、少数の企業によって支配されている。
- エンジニアリング標準とオープンソースの評価は、実践を民主化するのに役立つ。

◆ **統一されたガイドラインの必要性：**
- ドメイン固有の実装を始める前に、最低限の統一されたガイドラインが必要。
- 特に、人命への危険を回避するため、最低限のガイドラインは必須である。

◆ **責任分担：**
- セーフティでない結果に対する責任は、基盤モデル作成者から、アプリケーション開発者、デプロイ者、インフラ提供者、エンドユーザーまで、パイプラインの上下に及ぶべきである。
- 交通機関で使用されるモデル法的枠組みは、有用なアナロジー（類推）となる可能性がある。
- 開発者に利害関係者のリストアップを要求することが提言された。

◆ **文化に基づくインセンティブの方法：**
- インセンティブと罰則の方法は文化によって異なり、東アジア諸国では実際の行動と被害を観察し、指導を行う方法が好まれる。
- これはヨーロッパの予防的アプローチや、米国の規制当局への依存とは対照的。

AI開発における標準化・責任分担・透明性の重要性

AISI Japan AI Safety Institute

AI開発におけるエンジニアリング標準の必要性と、セーフティ評価の改善に関する重要なポイントを強調
AI開発における責任分担、文化的なインセンティブの違い、フロンティアモデル企業の透明性を強調

- **AIセーフティ範囲の考慮：**
  - 意図しない危害（バグや事故）と意図的な有害使用（自律型兵器など）の両方が、AIセーフティの範囲内で考慮されるべき。
  - ユーザーエラーや誤用だけでなく、比例加重責任としてはフロンティアモデル企業が最大の責任を負う。
- **フロンティアモデル企業の透明性：**
  - フロンティアモデル企業は、その活動やサプライチェーンにおいてもっとオープンにすべきであり、リスクは開発の初期段階で発見し、軽減する必要がある。
- **GPAI製品を使用する企業のリスク：**
  - GPAI製品の上にアプリケーションを構築する川下企業は、大きな依存リスクを抱えている。
  - これらの企業は関係するリスク／責任レベルを見積もり、列挙する必要がある。
- **評価と透明性の重要性：**
  - GPAIモデルの導入を検討している企業は、評価が採用の意思決定に与える影響について言及した。
  - フロンティアモデル企業によるセーフティの懸念、評価、緩和策についての透明性が重要。

AIリテラシー向上とセーフティ緩和策の実行

AISI Japan AI Safety Institute

AIリテラシーの向上と利害関係者の協力、社会技術的な背景の重要性に焦点
AIセーフティに対する懸念とその解決策の実行に関する課題を強調

◆ AIリテラシーの不足：

- 開発者とユーザーのAIリテラシーの不足が重要な課題として浮き彫りになった。
- 義務的リテラシー・プログラムや認証制度、広範な利害関係者の参加を求める声が繰り返し上がった。

◆ 情報提供と訓練の重要性：

- 開発者、実装者、ユーザーがより多くの情報に触れ、訓練を受けることで、セーフティの問題を検出し、ユースケースを提示し、採用を推進することに関与できる。

◆ 緩和策における利害関係者の参加：

- 緩和策には技術者だけでなく、市民社会、規制当局、エンドユーザーコミュニティの利害関係者も参加する必要がある。
- オープンな情報共有、参加型の標準開発、公的説明責任が繰り返し強調された。

◆ 社会技術的背景の重要性：

- 技術的な手法だけでは不十分であり、社会技術的な背景があらゆる解決策の適用を導く必要がある。

◆ AIセーフティの懸念と解決策：

- AIセーフティに関する懸念と可能な解決策は明確であるにもかかわらず、実際にはAIセーフティ対策を効果的に実施するための多くの課題がある。
- ワークショップの回答は、これらの課題を反映している。

AI技術進化に伴うセーフティ概念の多様化とリスク評価の不確実性

AISI Japan AI Safety Institute

AIセーフティの概念の多様性とその評価に関する不確実性、リスクの許容範囲についての議論を反映
AI技術の進化とそのセーフティ緩和策の適用における遅れを強調

◆ 「セーフティ」の多様な解釈：

- 「セーフティ」という概念が、現在、多くの人々にとってさまざまな意味を持つことが指摘された。
- この多様な解釈は、実務者に混乱をもたらし、AIシステムのセーフティの評価や管理を難しくしている。

◆ 不確実性と「十分なセーフティ」：

- 何が「十分なセーフティ」や「適切な予算」を構成するかについては、不確実性が残っている。
- いくつかの能力（再帰的自己改良、生物兵器の開発、自律的自己複製など）は、大惨事の可能性があるため、非合法とすべきだとの意見があったが、それをどう体系的に防ぐかについては意見が一致しなかった。

◆ リスクの許容範囲と法律：

- 法律は、許容できるリスクのレベルと許容できないリスクの定義を決定する必要がある。
- リスクをゼロにすることは不可能だが、何が容認できないかについて明確な定義と合意が必要だと参加者は指摘した。

◆ AIの進歩とセーフティ緩和策の課題：

- AI、特に基盤モデルの進歩の速さが、AIのセーフティ対策を適切に適用する際の重大な課題として再度言及された。
- 規制、安全、研究の各コミュニティは進歩に歩調を合わせることができていない。

◆ リスクのキャッチアップゲーム：

- リスクは常にキャッチアップゲームを繰り広げ、セーフティ対策が有意義に適用されるよりも早く進化する。
- 技術の進歩に追いつくのは当然だが、そのギャップは拡大している。

AIセーフティ評価の課題：手作業・透明性技術・評価方法・予算制約

AISI Japan AI Safety Institute

手作業の評価、透明性技術の採用制限、表面的なセーフティ対策、評価方法の限界を強調
AIセーフティに関連する予算の不足と、その改善のための提案、また評価方法に関するリスクを強調

◆ **手作業によるセーフティ評価の課題：**
- 手作業によるセーフティ評価は高価で時間がかかり、業界関係者に独占されがちである。
- 特に小規模な組織は、有意義な参加のための予算や専門知識が不足している。

◆ **透明性技術の採用範囲の狭さ：**
- 電子透かしやメタデータ署名のような透明性技術は、大手企業に限られており、その採用範囲は狭い。

◆ **「チェックボックスにチェックを入れる」努力の危険性：**
- セーフティを形式的に確認するだけでは、実効性に欠け、コンプライアンスも十分に機能しない可能性

◆ **評価方法の限界：**
- 発表者と参加者は、多くの評価方法がまだ予備的で、強力な科学的根拠を欠いていると強調した。

◆ **金銭的・組織的インセンティブとセーフティ不整合：**
- 参加者は、AIの研究予算におけるセーフティへの割り当てがわずか1%程度であり、航空や原子力のようなリスクの高い分野への投資（最大90%のリソースがセーフティとコンプライアンスに集中）とは大きな差があることを強調した。
- セーフティ、認証、継続的な監査、義務教育に特定の予算パーセンテージを割く必要がある。

◆ **最低の予算閾値の提案：**
- 参加者は、セーフティに関する予算として10～20%を最低の閾値とすべきだと主張した。
- セーフティのための予算には、ロバストなテストと検証を行うための計算リソースも含まれるべき。

◆ **評価ベンチマークの過剰適合のリスク：**
- 実際のセーフティよりもマーケティングのための評価ベンチマークに過剰に適合させるリスクが指摘された。

ブラックボックス型AIモデルと国際協力：リスクマネジメント・連携不足・地域的アプローチの違い

AISI Japan AI Safety Institute

ブラックボックス型モデルの問題点、企業間の連携不足、リスクマネジメントの重要性を強調
地域間でのアプローチの違いや、協力の重要性を強調

◆ **ブラックボックスモデルの課題：**
- 不透明な「ブラックボックス」型のAIモデルは、外部からの問い合わせや説明責任を困難にする。

◆ **産業界と研究コミュニティの連携不足：**
- 産業界と研究コミュニティ（市民社会を含む）との間の連携が不十分で、ガバナンス・モデルの深さと妥当性が制限される。

◆ **大手テクノロジー企業の偏り：**
- 大手テクノロジー企業は経済的または戦略的な目標に偏り、セーフティ対策を軽視したり遅らせたりすることがある。
- これにより、技術的負債が生まれ、川下の企業、個人、地域社会にさらなる負担とリスクがもたらされる。

◆ **リスクの排除と透明性の必要性：**
- 特定のリスクは排除できないが、リスク許容度を適切に評価し、対策を講じるためには、データ、モデル、設計上の意思決定に関する最低限の透明性が必要である。

◆ **AIセーフティに関する国際的な連携の重要性：**
- AIセーフティに関する取り組みが連携して行われるには、文化や政治体制、経済状況など地域ごとの違いを考慮しなければならない。
- 各国や地域の違いが、「セーフティ」の定義や優先順位に影響を与える可能性がある。

◆ **アジア諸国と欧米諸国のアプローチの違い：**
- アジア諸国では、社会的実施、変革、信頼性、ガバナンス・フレームワークの開発に重点が置かれている。
- ヨーロッパや北米では、公共の透明性、説明責任、解釈可能性が強調されることが多い。

◆ **社会技術的視点と解釈可能性の重要性：**
- 解釈可能性や社会的正当性が強調される分野（金融、教育、医療など）では、ヨーロッパやアジアで見られるアカウンタビリティ重視のアプローチと高い互換性があると考えられている。

◆ **国際協力の重要性：**
- 組織や国境を越えた協力が成功のための重要な要素であることに参加者は同意した。

AIセーフティガバナンスにおける柔軟性と社会技術的要素：分野ごとの適応とセーフティ向上の具体的ポイント

AISI Japan AI Safety Institute

AIセーフティガバナンスにおける柔軟性、社会技術的側面、分野ごとの適応に関する重要ポイントを強調
AIセーフティ向上のための具体的なポイントに焦点

◆ **AIセーフティのガバナンスにおける柔軟性の必要性：**
- AIセーフティについては広範な合意があるが、ガバナンスは制度的、文化的、社会的な差異を考慮した柔軟性を持つべきである。
- これは、地域の価値観を尊重し、画一的な解決策を押し付けないことを意味する。

◆ **社会技術的側面の重要性：**
- 地域コミュニティの目から見てシステムが正当なものであることを保証する社会技術的側面が、牽引力と信頼を築く上で極めて重要である。

◆ **金融、医療、教育分野でのバランス：**
- 特に影響力の大きい金融、医療、教育といった分野では、共通のコア（基本的権利、基本的安全基準など）と現地適応のバランスを取る必要がある。

◆ **連携強化のためのポイント：**
- **透明性のある標準化**：セーフティクリティカルなモデルで誰が作業しているかを透明化。
- **AI開発におけるパワーと能力の追跡**：人材と計算能力の集中を追跡。
- **AIのセーフティ評価ツールと監査インフラの構築**：これらを構築し、民主化する。
- **AIインシデントリポジトリの展開**：関係者全体で、既知の問題を繰り返さないようにし、より安全なアプローチに貢献していく。
- **責任の分散**：開発と展開のパイプライン全体で責任を分散する。
- **AIリテラシーと認識の向上**：利害関係者の間でAIに対するリテラシーと認識を高め、職場の慣行や公共政策にトレーニングを組み込む。

本日のアジェンダ

AISI Japan AI Safety Institute

1. AIセーフティ・インスティテュートの紹介とその役割
2. AIセーフティ評価を巡る最新動向
    - GPAI Tokyo Innovation Workshop
    - Japan AI Safety Workshop
    - 国連専門機関ITU主催 AI for Good Global Summit 2025
    - Berkeley Agentic AI Summit 2025
    - Int'l Conference on Auditing and Artificial Intelligence
3. 品質マネジメントの重要性

国連専門機関ITU主催 AI for Good Global Summit 2025
（2025年7月、スイス・ジュネーブ　169か国・11,000名参加）

AISI Japan AI Safety Institute

単一テストを超えた、多層的・共同的な検証エコシステムの構築

【AI評価-技術指標と社会的価値のギャップ】

- **性能評価の視点の拡張**

  従来：モデルの性能指標（精度・再現率など）の確認に偏重

  課題：前提条件や概念の妥当性が見落とされがち

  必要な視点：

  社会的コンテキスト・利用環境を考慮

  コミュニティの成熟度や受容性を反映

  実社会での有効性・信頼性を検証

- **単なる数値的性能評価から「実社会で機能するか」の総合的評価へ**

【適合性評価の進化】

- **AIセーフティと倫理は不可分**

  前提：AIセーフティは倫理的視点と切り離せない

- **評価の焦点**：
  - 技術的側面：セーフティ・信頼性の確保
  - 組織的側面：価値観や概念がマネジメントに組み込まれているか

- **重要性**：両面からの統合的な評価が、責任あるAI活用につながる

(Source: https://aiforgood.itu.int/speaker/hiromu-kitamura/　J-AISI北村プレゼン資料から抜粋)

国連専門機関ITU主催 AI for Good Global Summit 2025
（2025年7月、スイス・ジュネーブ　169か国・11,000名参加）

AISI Japan AI Safety Institute

単一テストを超えた、多層的・共同的な検証エコシステムの構築

【ベストプラクティスの共有】

- **未知のリスクへの対応**

  課題：経験則だけでは新しい・複雑なリスクを十分に扱えない

- **必要なアプローチ**：
  - 論理的な枠組みをエコシステム全体で共有
  - 「リスクチェーンモデル」を導入し、システム間の相互作用を把握

- **狙い**：個別最適ではなく、全体最適でリスクを捉え、予防と対応力を強化

【AI テストの課題】

- **テストの再定義の必要性**

  背景：AIの社会的影響が拡大

  課題：「何をテストするか」だけでは不十分

- **新たな問い**：「どこまでテストすべきか」

方向性：
  - 実世界での影響を踏まえた検証
  - テストの範囲と深さを再定義

(Source: https://aiforgood.itu.int/speaker/hiromu-kitamura/　J-AISI北村プレゼン資料から抜粋)

国連専門機関ITU主催 AI for Good Global Summit 2025
（2025年7月、スイス・ジュネーブ　169か国・11,000名参加）

AISI Japan AI Safety Institute

単一テストを超えた、多層的・共同的な検証エコシステムの構築

【国際協力の必要性】

- **グローバル基盤の必要性**

  背景：技術的・社会技術的な変曲点に直面

  多様なリスクシナリオ：

  - パンデミック型リスク
  - 悪意ある攻撃者によるリスク
  - 無害な事象の同時発生による複合リスク

  課題：画一的アプローチでは対応できない

  必要性：各国がリスクパターンと対応策を 越境的に共有 し、協力的なグローバル基盤を構築

(Source: https://aiforgood.itu.int/speaker/hiromu-kitamura/　J-AISI北村プレゼン資料から抜粋)

本日のアジェンダ

AISI Japan AI Safety Institute

1. AIセーフティ・インスティテュートの紹介とその役割
2. AIセーフティ評価を巡る最新動向
    - GPAI Tokyo Innovation Workshop
    - Japan AI Safety Workshop
    - 国連専門機関ITU主催 AI for Good Global Summit 2025
    - Berkeley Agentic AI Summit 2025
    - Int'l Conference on Auditing and Artificial Intelligence
3. 品質マネジメントの重要性

Berkeley Agentic AI Summit 2025
（2025年8月、米国・バークレー）

AISI Japan AI Safety Institute

Agentic AIによる新価値創出と実証競争の加速

医療（Healthcare）
- 患者データの抽出と管理
- 医療現場での患者とのコミュニケーション
- 臨床試験のサポート
- 新薬の開発や治療法の提案

農業（Agriculture）
- 作物の管理や生育の監視
- 農業の生産性向上に向けたAI活用
- 天候データや土壌情報を基にした農業支援

製造（Manufacturing）
- 自動化された品質管理
- サプライチェーンの最適化
- メンテナンス予測とエラー検出

金融（Finance）
- リスク評価や不正取引検出
- 顧客サポートの自動化
- 投資判断支援

カスタマーサポート（Customer Support）
- AIエージェントによる顧客対応
- チャットボットを使った24時間対応

教育（Education）
- 学習支援のためのパーソナライズドAI
- 自動化された試験の採点やフィードバック

(Source: Center for Responsible, Decentralized Intelligence at Berkeley)

Berkeley Agentic AI Summit 2025（2025年8月、米国・バークレー）

AISI Japan AI Safety Institute

技術革新から社会的インパクトへ ― Agentic AIの展望

【発展の兆し】

◆ Agentic AIの可能性

役割の進化：業務効率化を超え、社会的価値創出や意思決定のパートナーへ

高度な能力：自己学習・自律的意思決定により、不確実な状況でも最適な判断

創造性の発揮：ビジネス戦略や研究開発で新しいアイデアを提案

社会課題への貢献：気候変動や貧困など、複雑課題への解決策提示

【技術進展とその影響】

◆ Agentic AIの進化と影響

- **急速な進化**：今後数年で社会への影響力が飛躍的に拡大
- **求められる対応**：倫理・責任の視点を含めた適切なガバナンス
- **技術的進展**
- 高精度な予測・最適化
- 複雑・動的な環境でも学習・意思決定可能
- 人間を補完し、一部では超える判断力
- **応用分野**
- **医療**：迅速かつ個別化された診断・治療、疾患予測と予防提案
- **金融**：市場解析、リスク管理、資産運用支援、投資戦略の高度化

(Source: Center for Responsible, Decentralized Intelligence at Berkeley)

本日のアジェンダ

AISI Japan AI Safety Institute

1. AIセーフティ・インスティテュートの紹介とその役割
2. AIセーフティ評価を巡る最新動向
    - GPAI Tokyo Innovation Workshop
    - Japan AI Safety Workshop
    - 国連専門機関ITU主催 AI for Good Global Summit 2025
    - Berkeley Agentic AI Summit 2025
    - Int'l Conference on Auditing and Artificial Intelligence
3. 品質マネジメントの重要性

Int'l Conference on Auditing and Artificial Intelligence（2025年8月、ドイツ・デュイスブルク）

AISI Japan AI Safety Institute

University of Duisburg-Essen in Duisburg, Germany主催で開催
監査分野における人工知能（AI）導入の可能性と課題を国際的に議論

- AIの監査応用
  技術的転換は、監査の定義や責任構造そのものに揺さぶり。
- AIの監査応用で、AIの開発・実装・運用の本質的な課題が改めて浮き彫りに。

(Source: 2nd International Conference on Auditing and Artificial Intelligence – 2025)

Int'l Conference on Auditing and Artificial Intelligence（2025年8月、ドイツ・デュイスブルク）

AISI Japan AI Safety Institute

浮かび上がった主要論点は大きく5つに分類

【効率化と有効性の両立】【法的責任とAI統治】【非構造データ処理とフレームワーク整備】【国際ガバナンスと規制整合】【人材と教育】

【効率化と有効性の両立】

監査部門の課題：少人数で膨大な対象をカバー
- **AIの可能性**：効率化に寄与するが、質低下のリスクも
- **事例**：Deutsche Bahnは「効率化25%」を目標 → 参加者からは「効率性より有効性を優先」との指摘
- **有効性の定義**：リスク見逃しの低減、意味ある監査意見の提供
- **結論**：AIは効率性と有効性の両立を支援し得るが、設計思想が鍵

【法的責任とAI統治】
- **未解決の課題**：AIの誤判定で重大な不備や不正を見逃した場合、責任の所在（監査法人・AI提供者・監査人個人）は不明確
- **現状**：国際的に統一された指針は存在しない
「AIはあくまで補助、最終判断は人間」
- **懸念**：実際にはAIへの依存度が高まり、「人間の責任」が形式的に残るだけとなる危険

(Source: 2nd International Conference on Auditing and Artificial Intelligence – 2025)

Int'l Conference on Auditing and Artificial Intelligence（2025年8月、ドイツ・デュイスブルク）

AISI Japan AI Safety Institute

浮かび上がった主要論点は大きく5つに分類

【効率化と有効性の両立】【法的責任とAI統治】【非構造データ処理とフレームワーク整備】【国際ガバナンスと規制整合】【人材と教育】

【非構造データ処理とフレームワーク整備】

- **課題**：ESG報告など非財務情報の監査は非構造データが中心
- **AIの挑戦**：データを構造化・分析可能に変換する新しい枠組み
- **事例**：DePaul大学の研究
  - 4エージェント（プロンプト生成→抽出→評価→ベンチマーク）が連携
- **懸念点**：
  - 意図や意味が誤解されるリスク
  - 規制当局がAI生成データを証拠として認めるのか不透明

【国際ガバナンスと規制整合】

現状：OECD・EU・米国・日本の政策は方向性は類似するが、細部は異なる

- **課題**：監査は国境を越えるため、規制の不整合が大きな障害に
- **動向**：EU当局者が「AI Actの監査分野への拡張可能性」に言及し、実務家の注目を集めた

(Source: 2nd International Conference on Auditing and Artificial Intelligence – 2025)

Int'l Conference on Auditing and Artificial Intelligence （2025年8月、ドイツ・デュイスブルク）

AISI Japan AI Safety Institute

浮かび上がった主要論点は大きく5つに分類

【効率化と有効性の両立】【法的責任とAI統治】【非構造データ処理とフレームワーク整備】【国際ガバナンスと規制整合】【人材と教育】

【人材と教育】

- **最大の課題**：人間側の準備不足
- **事例**：Volkswagen「新ツール定着に時間」
- **対応策の議論**：教育・マイクロクレデンシャル制度の整備
- **指摘**：J-AISI「監査人のAIリテラシー向上が必要」

(Source: 2nd International Conference on Auditing and Artificial Intelligence – 2025)

本日のアジェンダ

AISI Japan AI Safety Institute

1. AIセーフティ・インスティテュートの紹介とその役割
2. AIセーフティ評価を巡る最新動向
    - GPAI Tokyo Innovation Workshop
    - Japan AI Safety Workshop
    - 国連専門機関ITU主催 AI for Good Global Summit 2025
    - Berkeley Agentic AI Summit 2025
    - Int'l Conference on Auditing and Artificial Intelligence
3. 品質マネジメントの重要性

品質マネジメントの重要性

AISI Japan AI Safety Institute

AI時代における基盤的能力と適応力の強化

◆ SQiP (Software Quality Profession)

実践的で実証的なソフトウェア品質技術・施策の研究・普及を目的として、日本科学技術連盟の下に設置されたソフトウェア品質向上のための活動

✓ ソフトウェア品質保証の極意
“日本科学技術連盟　ソフトウェア品質保証プロフェッショナルの会”

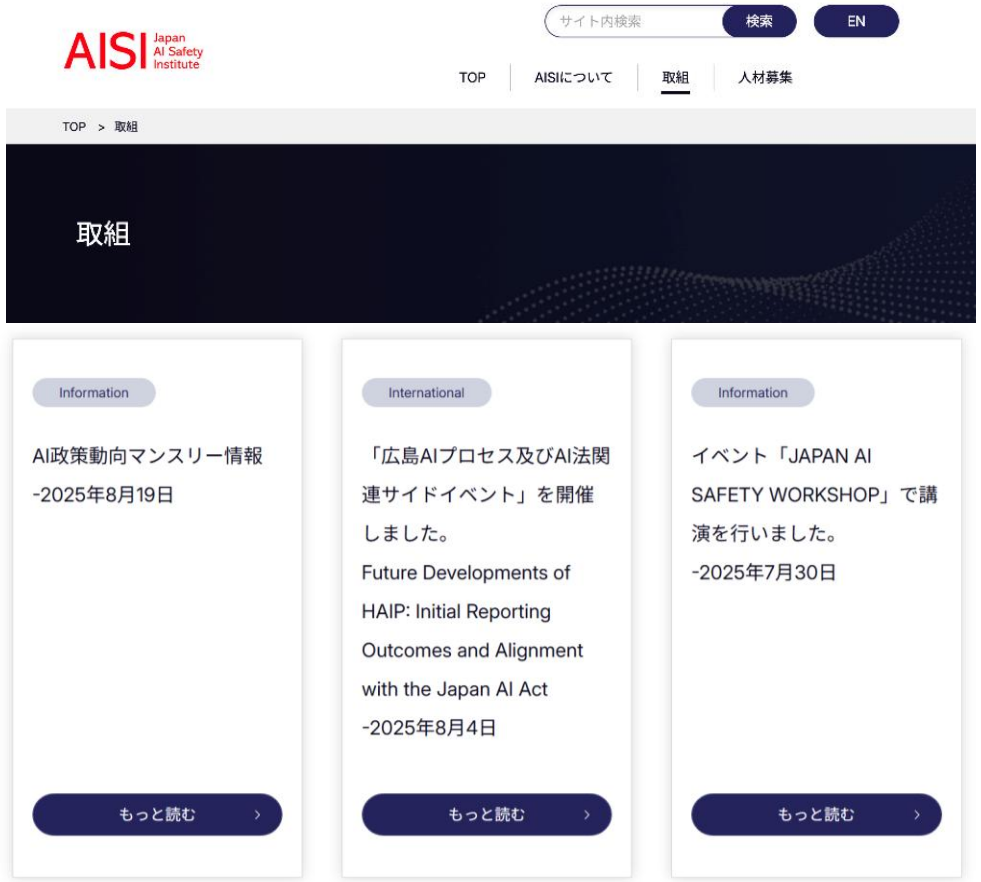
(Source : https://www.juse.or.jp/sqip/community/bucyo/index.html)

- **“Ignite Your Insight, Unlock Your Potential,”**
  “Adapt with Awareness. Engage with Humility. Grow with Purpose.”

(Source : Katsue Kojima -https://www.engineer.or.jp/c_cmt/danjyo/topics/010/attached/attach_10299_9.pdf)

まずはAISIの活動に注目！

AISIではホームページやLinkedInで逐次情報を展開しています
AISI Japan - AI Safety Institute



AISI

Japan AI Safety Institute